Very preliminary draft 2/16/26
All rights reserved

LIBERAL AI

Cass R. Sunstein*

Abstract

Can AI be liberal? In what sense? One answer points to the liberal insistence on freedom of choice, understood as a product of the commitment to personal autonomy and individual dignity. Mill and Hayek are of course defining figures here, emphasizing the epistemic foundations for freedom of choice. "Choice Engines," powered by AI and authorized or required by law, might promote liberal goals (and in the process, produce significant increases in human welfare). A key reason is that they can simultaneously (1) preserve autonomy, (2) respect dignity, and (3) help people to overcome inadequate information and behavioral biases, which can produce internalities, understood as costs that people impose on their future selves, and also externalities, understood as costs that people impose on others. Different consumers care about different things, of course, which is a reason to insist on a high degree of freedom of choice, even in the presence of internalities and externalities. AI-powered Choice Engines can respect that freedom, not least through personalization. Nonetheless, AI-powered Choice Engines might be enlisted by insufficiently informed or self-interested actors, who might exploit inadequate information or behavioral biases, and thus co5mpromise liberal goals. AI-powered Choice Engines might also be deceptive or manipulative, again compromising liberal goals, and legal safeguards are necessary to reduce the relevant risks. Illiberal or antiliberal AI is not merely imaginable; it is in place. Still, liberal AI is not an oxymoron. It could make life less nasty, less brutish, less short, and less hard—and more free.

I. CHOICE AND WELFARE

Can AI be liberal? In what sense? One answer points to the liberal insistence on freedom of choice, understood as a product of commitments to personal autonomy and individual dignity. AI can respect those commitments. We can imagine liberal AI, and we can imagine illiberal or antiliberal AI.

* Robert Walmsley University Professor, Harvard University. This paper is an early draft of the Ethics in AI Lecture, to be delivered at Jesus College, Oxford University on May 22, 2026. Some parts of the paper draw on previous, more applied work on Choice Engines. I am grateful to the Harvard Law School's Program on Behavioral Economics and Public Policy, and the Harvard Law School's Initiative on Artificial Intelligence and Law, for valuable support.

In *On Liberty*, Mill insisted[1]:

> The only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will make him happier, because, in the opinion of others, to do so would be wise, or even right.

In the first instance, Mill's argument for his famous Harm Principle is epistemic. Mill believed that choosers are in the best position to know what is good for them. In Mill's view, the problem with outsiders, including government officials, is that they lack the necessary information. Mill contended that the individual "is the person most interested in his own well-being," and the "ordinary man or woman has means of knowledge immeasurably surpassing those that can be possessed by anyone else." When society seeks to overrule the individual's judgment, it does so on the basis of "general presumptions," and these "may be altogether wrong, and even if right, are as likely as not to be misapplied to individual cases." If the goal is to ensure that people's lives go well, Mill concludes that the best solution is to allow people to find their own path. (On behalf of the Harm Principle, Mill offered romantic[2] and vaguely Kantian arguments as well as epistemic ones. We might even say that the words of *On Liberty* are epistemic, but the music is romantic.)

Also focused on coercion, Friedrich Hayek too showed little enthusiasm for paternalism, and while he did not discuss it as such, his work on liberalism and liberty can easily be read to stand against it: "Coercion is evil precisely because it thus eliminates an individual as a thinking and valuing person and makes him a bare tool in the achievement of the ends of another."[3] This is a strikingly Kantian formulation, not speaking of welfare at all; notice the use of the word "evil" and the objection to treating people as means rather than ends. And indeed, Hayek seemed to embrace something like a Kantian, nonwelfarist foundation for freedom. In his introduction to *The Constitution of Liberty*, he wrote, "Some readers will perhaps be disturbed by the impression that I do not take the value of individual liberty as an indisputable ethical presupposition and that, in trying to demonstrate its value, I am possibly making the argument in its support a matter of expediency. That would be a misunderstanding."[4] But at pivotal points, Hayek extended the

---

[1] JOHN STUART MILL, ON LIBERTY 80 (David Bromwich & George Kateb eds., 2003) (1859).
[2] On Mill's romantic streak, see Cass R. Sunstein, On Liberalism (2025).
[3] Friedrich Hayek, The Constitution of Liberty 71 (Ronald Hamowy ed. 2011).
[4] Friedrich Hayek, The Constitution of Liberty 52 (Ronald Hamowy ed. 2011).

epistemic argument in favor of markets to make a general epistemic claim of his own, not at all rooted in "an indisputable ethical presupposition," on behalf of liberty and against coercion. Consider his suggestion that 'the awareness of our irremediable ignorance of most of what is known to somebody [the chooser] is the chief basis of the argument for liberty. This is especially true in the economic field. If it appears that the market mechanism leads to the effective utilization of more knowledge than any directing agency can possess, this is the chief foundation of the case for economic freedom …." That passage is in a long-unpublished lecture. He put a similar point more concisely in *The Constitution of Liberty*: "The case for individual freedom rests chiefly on the recognition of the inevitable ignorance of all of us concerning a great many of the factors on which the achievement of our ends and welfare depends."[5]

Does AI undermine the case for individual freedom? Does it throw Mill's Harm Principle in question? When do people's choices in fact promote their autonomy? When do such choices make their lives go better? When do external interventions, designed to improve people's lives, actually help to increase autonomy and welfare? How should private and public institutions, and all those involved in policy and law, proceed if populations are diverse, and if some people will be affected differently from others? How might AI help or hurt? My purpose here is to make some progress in answering these questions, above all by showing that liberal AI is no oxymoron, and that AI, committed to liberal principles, can help people to overcome an absence of information and behavioral biases. It is emphatically true that AI may or may not be liberal, and also that it may or may not paternalistic. In some cases, it might turn out to engage in libertarian paternalism, preserving freedom of choice while steering people in directions that make their lives go better.[6] Paternalistic AI might significantly improve people's welfare, and the dangers to autonomy, and to liberal commitments, are reduced if freedom of choice is maintained. At the same time, AI poses serious risks, including risks of deception and manipulation; legal safeguards will be required.

II. The (Relevant) Perils of Average Treatment Effects

[5] Id. at 80.

[6] *See generally* RICHARD THALER & CASS R. SUNSTEIN, NUDGE: THE FINAL EDITION (2021). Earlier accounts, much developed and expanded upon here, can be found in Cass R. Sunstein, Choice Engines and Paternalistic AI, 11 Humanities and Social Sciences Communications, article number 888 (2024), and Cass R. Sunstein, Brave New World? Human Welfare and Paternalistic AI, 26 Theoretical Inquiries in Law 1 (2025). At times I draw on those accounts here.

To approach these questions, and to see the complex relationships between freedom, choice, and paternalism, consider three sets of findings:

1. Calorie labels, required by law, are designed to help people make better choices about what to eat, and there is evidence that calorie labels have welfare benefits by doing exactly that.[7] At the same time, calorie labels seem to have a greater effect on people who do not have self-control problems than on people who suffer from such problems.[8] That is a serious problem, because a central point of calorie labels is liberal in character: to help people with self-control problems. It is possible that in some populations, calorie labels affect people who do not need help and have little or no effect on people who do need help, except to make them feel sad and ashamed. If so, calorie labels could make people worse off.
2. On average, people appear to benefit from home energy reports, which inform them about their energy usage. As a result of such reports, people save money, on average. They are willing to pay, on average, a positive amount to receive such reports. But some people are willing to pay far more than others for home energy reports.[9] In fact, some people are willing to pay *not* to receive home energy reports. They dislike receiving such reports, and they believe that they would be better off without them. While home energy reports are designed to save consumers money and to reduce externalities, sending such reports to some people, who do not want to receive them, seems to make them worse off. More targeted use of home energy reports, sending them only to those who want and would benefit from them, could produce significant welfare gains.[10] More targeted use of such reports would seem to fit better with liberal principles as well, because they respect freedom of choice.

---

[7] Linda Thunström, *Welfare Effects of Nudges: The Emotional Tax of Calorie Menu Labeling*, 14 Judgment & Decision Making 11, 18-19 (finding that a substantial number of study participants favored calorie labels because "calorie content would matter to my meal choice," *id.* at 19) (2019).
[8] *Id.* at 20.
[9] Hunt Allcott & Judd Kessler, *The Welfare Effects of Nudges*, 11 AM. ECON. J.: APPLIED ECON. 236, 257 (2019).
[10] *See id.* at 271 (proposing "an opt-out program that targets consumers who would generate large welfare gains and excludes consumers who would not").

3. Graphic warning labels on sugary drinks are designed to affect consumer behavior, and there is evidence that they can do so.[11] But such labels have greater effects on some consumers than on others.[12] Disturbingly, such labels can lead people who do not have self-control problems to consume less in the way of sugary drinks, while having a significantly smaller effect on people who do have self-control problems. In addition, many people do not like seeing graphic warning labels. The average person in a large sample reported being willing to pay about $1 to *avoid* seeing the graphic warning labels, which would seem to be strong evidence that seeing such labels would reduce their welfare.[13] It is likely that graphic labels are helping some and hurting others. It is possible that such labels are on balance causing harm, in the sense that they are reducing people's welfare in aggregate.

From these sets of findings, we can draw three simple conclusions. *First,* legal interventions, designed to promote freedom of choice and to improve people's lives, might well have different effects on different populations. Under favorable conditions, they might have large positive effects on people who need help, and small or no effects on people who do not need help. Under unfavorable conditions, they might have small or no effects on a group that needs help, and large effects on a group that does not need help. Large effects on a group that does not need help may not improve that group's welfare. For example, people who have no need or reason to change their spending patterns or their diets might end up doing so.

Second, legal interventions may have either positive or negative *hedonic* effects. Liberals care about those effects, which are central to the liberal commitment to freedom of choice. People might like seeing labels, or they might dislike seeing labels. They might like receiving information, or they might not like receiving information. With respect to the receipt of information, there is a great deal of heterogeneity.

Third, and consistent with the first conclusion, an understanding of the average treatment effect of an intervention does not tell us about the overall effect on liberty and social welfare, which is

[11] Hunt Allcott et al., *When Do "Nudges" Increase Welfare*? 4 (Nat'l Bureau of Econ. Rsch., Working Paper No. 30740, 2022), https://www.nber.org/system/files/working_papers/w30740/w30740.pdf.
[12] *Id.*
[13] *Id.* at 25.

what regulators and others involved in law and policy need to know.[14] The third conclusion is easy to miss. An intervention might have a positive average treatment effect, leading to healthier choices, but if people who make healthier choices are already healthy, and if people who do not make healthier choices are sad and ashamed, then the interventions might turn out to reduce social welfare. If a legal intervention makes people feel sad and ashamed, it is, to that extent, reducing people's welfare. It is not a lot of fun to feel sad or ashamed.

These points can be made about a wide range of legal interventions, increasingly adopted with reference to AI. They hold for taxes that are designed to produce more informed and healthier choices: it is possible that such taxes will have little or no effect on the people they are particularly intended to help, while having a significant adverse effect on people they are not (particularly) intended to help. (Consider soda taxes.) They hold for mandates, understood as requirements backed by the force of law, and also for bans, understood as prohibitions similarly backed. A ban on some activity or product might, on balance, hurt people who greatly benefit from that activity or product, while helping people who lose only modestly from it. (Consider bans on the purchase of incandescent lightbulbs.) In all of these cases, more targeted action, treating different individuals and groups differently, would be better than "mass" action, treating everyone the same. They hold for automatic enrollment: it is possible that automatic enrollment in some plan will have no effects on people who benefit from enrollment (because they would enroll in any case) while harming people who do not benefit from enrollment (because some or many who lose do not opt-out, perhaps because of inertia).

Here, then, is where we are. Interventions designed to influence people's choices may or may not fit with the liberal commitment to freedom of choice (and may or may not increase social welfare). A recurring problem is that of heterogeneity. Following Mill and Hayek, contemporary liberals insist that people have different needs, preferences, and values, and an intervention that affects a large population might help some and hurt others. The average treatment effect (that is, the effect of the treatment on the average person) is not the same as the effects on liberty and welfare. Those effects are what matters. How can they be improved? AI, committed to liberal

[14] *See id.* at 29 ("While much of the empirical literature has focused on whether nudges have average effects in the 'right' direction, we show that welfare also depends on how nudges affect the variance of distortions.").

strictures and meant to improve social welfare, can provide an answer, at least if it is focused on (1) harms that people do to their future selves and (2) harms that people do to others.

III. AI-POWERED CHOICE ENGINES, LIBERAL AND ILLIBERAL

AI-powered Choice Engines[15] can be defined as instruments, typically (but not always) online, by which choosers are given an opportunity to provide some information about themselves and their preferences, and then receive a recommendation, or a set of recommendations, about what they ought to select. I am aware that the term and the concept might seem a bit ominous, not least from the liberal point of view, but let us put that to one side. We will return to the matter of ominousness.

Begin with consumer behavior, and suppose that you are choosing a new laptop. You might say that you want to pay less than $1200; that you want the laptop to weigh less than three pounds; and that you want the laptop to have at least eight hours of battery life. A Choice Engine might give you a set of options. It will increasingly use AI. It might well be an algorithm: people with characteristics and preferences A, B, and C typically prefer, in order, laptops X, Y, and Z. Any kind of AI-powered Choice Engine might serve to specify and narrow your choices. Choice engines of this kind might be used for any number of choices, including (to take some random examples) choices of dogs,[16] mystery novels, cellphones, shavers, romantic partners, shoes, tennis racquets, ties, investments, political candidates, and political positions. Choice engines may or may not use AI, and if they do, they can use AI of different kinds. They may or may not be liberal, in the particular sense of being committed to freedom of choice.

One of my concerns here is how liberal Choice Engines, powered by AI and permitted, regulated, or required by law, might address an absence of information or a behavioral bias, and

[15] The term is not in general use, but something like it can be found in various places, with variations. *See* Michael Yeomans et al., *Making Sense of Recommendations*, 32 J. BEHAV. DECISION MAKING 403, 403 (2019); Guy Champnis, *The Rise of the Choice Engine*, ENERVEE (Mar. 6, 2018, 12:00 AM), https://www.enervee.com/blog/the-rise-of-the-choice-engine; *Why Getting Help Matters,* EDELMAN FINANCIAL ENGINES http://corp.financialengines.com/individuals/why-getting-help-matters.html (last visited July 22, 2024). Compare Choice Engines to the following, which is regrettably complicated: *Buying a Refrigerator Guide: How to Choose a New Fridge in 2024*, WHIRLPOOL, https://www.whirlpool.com/blog/kitchen/buying-guide-refrigerator.html (last visited July 22, 2024).

[16] *See Welcome to the Purina Dog Breed Selector*, PURINA, https://www.purina.co.uk/find-a-pet/dog-breeds/breed-selector (last visited July 22, 2024).

thus promote autonomy and increase social welfare. I will be drawing attention to the importance of targeting and personalization and the promise of AI-powered Choice Engines in improving individual choices. This will be a largely optimistic account, focused on the extraordinary promise of Choice Engines, but I will also have something to say about threats, including the risk of deception and manipulation. If more space will be given to reasons for optimism than to threats, and the risk of illiberalism, it is not because threats and risks do not deserve attention; they are exceedingly important and should be kept in mind throughout.

Turn to a simple example. For retirement plans, many U.S. employers use something like a Choice Engine, designed to respect employee autonomy and broadly speaking, to comport with liberal principles. Employers know a few things about their employees (and possibly more than a few). On the basis of what they know, they automatically enroll their employees in a specific plan, suitable for their circumstances – the kind of plan that employees would choose if they were adequately informed and free from behavioral biases. The plan is frequently a diversified, passively managed index fund. Employees can opt out and choose a different plan, if they like. Alternatively, employers might offer employees a specified set of options, with the understanding that all of them are suitable, or suitable enough. (Options that are not suitable should not be included.) They might provide employees with simple information by which to choose among them. The options might be identified or rethought with the assistance of AI or some kind of algorithm.[17]

Here is one reasonable approach, consistent with liberal principles: Automatically enroll employees in a plan that is most likely to improve their well-being, given everything relevant that is known about them.[18] That is the plan that they would choose, if they were adequately informed and free from behavioral biases. What suits people in their thirties is not the same as what suits people in their sixties. To be sure, identification of that plan might prove daunting; it might be challenging to know what plan really is the right one for (say) a risk-averse fifty-year-old male, with two children, earning $150,000 annually. But a large number of plans can at least be ruled out.[19] Note that if the focus is on improving employee well-being, we are not necessarily speaking

[17] *See, e.g.*, FIDELITY GO https://www.fidelity.com/managed-accounts/fidelity-go/overview (last visited July 22, 2024).
[18] This is consistent with IAN AYRES & QUINN CURTIS, RETIREMENT GUARDRAILS 159-61 (2023).
[19] *See id*. at 160.

of revealed preferences. People might choose plans without sufficiently appreciating the opportunities or the risks.

For retirement savings, we can easily imagine many different kinds of Choice Engines. Some of them might be mischievous, in the sense that they are poorly suited to workers' situations; some of them might be fiendish and illiberal, in the sense that they help the plan's provider, not the workers; some of them might be random; some of them might be coarse or clueless; some of them might show behavioral or other biases of their own; some of them might be self-serving.[20] For example, people might be automatically enrolled in plans with high fees. They might be automatically enrolled in plans that are not diversified. They might be automatically enrolled in money market accounts, which offer low returns. They might be automatically enrolled in dominated plans.[21] They might be automatically enrolled in plans that are especially ill-suited to their situations. They might be given a large number of options and asked to choose among them, with little relevant information, or with information that leads them to poor choices.

The general point is that in principle, Choice Engines might respect liberal commitments and work to overcome behavioral biases.[22] For retirement plans, Choice Engines may or may not be paternalistic. If they are not paternalistic, it might be because they simply provide a menu of options, given what is known about relevant choosers.[23] If they are paternalistic, they might be mildly paternalistic, moderately paternalistic, or highly paternalistic. A moderately paternalistic Choice Engine might impose nontrivial barriers to those who seek certain kinds of plans (such as those with high fees). Consistent with liberal principles, the barriers might take the form of information provision, "are you sure you want to?" queries, and requirements of multiple clicks. We might think of a moderately paternalistic Choice Engine as offering "light patterns," understood as strategies that help consumers, as contrasted with "dark patterns," understood as strategies that harm consumers (for example, by automatically enrolling them in costly programs that have little or no value).[24] A highly paternalistic Choice Engine, putting some pressure on

[20] An episode of Black Mirror could easily be based on such scenarios. *Black Mirror* (Netflix).
[21] *See* AYRES & CURTIS, *supra* note 18.
[22] *See* Vidya Athota et al., *Can Artificial Intelligence (AI) Manage Behavioural Biases Among Financial Planners?*, 31 J. GLOB. INFO. MGMT. 1, 7-9 (2023). For a disturbing set of findings, *see* Yang Chen et al., *A Manager and an AI Walk into a Bar: Does ChatGPT Make Biased Decisions Like We Do?* (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4380365.
[23] *See Welcome to the Purina Dog Breed Selector*, *supra* note 16.
[24] S*ee generally* Jamie Luguri & Lior Strahilevitz, *Shining A Light on Dark Patterns*, 13 J. LEGAL ANALYSIS 43 (2021).

liberal principles, might forbid employees from selecting any plan other than the one that it deems in the interest of choosers, or might make it exceedingly difficult for employees to do that.

Now broaden the viewscreen and turn to questions outside of the domain of retirement plans: Would you like to buy an energy-efficient refrigerator that would cost you $X today, but save you ten times $X over the next ten years? What kind of car would you like to buy? Would you like to buy a fuel-efficient car that would cost you $800 more upfront than the alternative, but that would save you $8000 over the next ten years? What characteristics of a car, or a refrigerator, matter most to you? Do you need a large car? Do you like hybrids? Are you excited about electric cars, or not so much?

A great deal of work finds that consumers suffer from "present bias."[25] Current costs and benefits loom large; future costs and benefits do not. For many of us, the short term is what matters most, and the long term is a foreign country. The future is Laterland, a country that we are not sure that we will ever visit. This is so with respect to choices that involve money, health, safety, and more. There are plausible evolutionary explanations for present bias. If you are running from a tiger, you ought not to spend much time thinking about long-term monetary gains (or your retirement savings). But under modern circumstances, present bias can get you into a great deal of trouble.

Liberal AI need not suffer from present bias.[26] Imagine that you are able and willing to consult AI to ask it what kind of car you should buy. Imagine too that you discover that you are, or might be, present-biased, in the sense that you prefer a car that is not (according to AI) the one that you should get. What then? We could easily imagine liberal Choice Engines for motor vehicle purchases in which different consumers provide relevant information about their practices, their preferences, and their values, and in which the relevant Choice Engine immediately provides a set of options—say, good, better, and best. Something like this could happen in minutes or even

[25] *See, e.g.*, Joachim Schleich et al., *A Large Scale Test of the Effects of Time Discounting*, 80 ENERGY ECON. 377 (2019); Madeline Werthschulte & Andreas Loschel, *On the Role of Present Bias and Biased Price Beliefs in Household Energy Consumption*, 109 J ENV. EC. AND MANAGEMENT (2021); Theresa Kuchler & Michaela Pagel, *Sticking to Your Plan: The Role of Present Bias for Credit Card Paydown,* 139 J. FIN. ECON. 359 (2018), https://www.nber.org/system/files/working_papers/w24881/w24881.pdf; Ted O'Donoghue & Matthew Rabin, *Present Bias: Lessons Learned and to Be Learned,* 105 AM ECON REV 273 (2015); Jess Benhabib et al., *Present Bias, Quasi-Hyperbolic Discounting, and Fixed Costs,* 69 GAMES ECON. BEHAV. 205 (2010); Yang Wang & Frank Sloan, *Present Bias and Health,* J. RISK UNCERTAINTY 177 (2018). Importantly, Wang and Sloan find strong evidence of present bias in connection with health-related decisions.

[26] It might. Yang Chen et al., *supra* note 22 (manuscript at 10).

seconds, perhaps a second or two. If there are three options—good, better, and best—verbal descriptions might explain the ranking. Or a Choice Engine might simply say: best for you. It might do so while allowing you to see other options if you indicate that you wish to do so. It may or may not be paternalistic or come with guardrails designed to protect consumers against serious mistakes.[27]

IV. INTERNALITIES AND EXTERNALITIES

Attempting to respond to the kinds of findings with which I began, liberal Choice Engines might well focus solely on particular choosers and what best fits their particular situations. They might ask, for example, about what particular consumers like most in cars, and they might take account of the full range of economic costs, including the costs of operating a vehicle over time. If so, Choice Engines might have a paternalistic feature insofar as they suggest that Car A is "best" for a particular consumer, even if that consumer would not give serious consideration to Car A. A Choice Engine would attempt to overcome both informational deficits and behavioral biases on the part of those who use them. Freedom of choice would be preserved, in recognition of the diversity of individual tastes, including preferences and values.

Present bias is, of course, just one reason why consumers might not make the right decisions, where "right" is understood by reference to their own welfare. (So much for Mill.) Consumers might also suffer from a simple absence of information; they might not know a lot about cars and their relevant features. Or they might suffer from another kind of behavioral bias: unrealistic optimism, availability bias (assessing risks by asking what readily comes to mind), loss aversion, status quo bias (showing an excessive preference for the status quo), or limited attention (attending to only a subset of relevant features of a product or activity). Consumers might be vulnerable to anchoring (showing an undue effect from an initial number) or to framing effects. If people are making their own lives worse for any of these reasons, Choice Engines might help. They might be paternalistic (and liberal) insofar as they respond to behavioral biases on the part of choosers, perhaps by offering recommendations or defaults, perhaps by imposing various barriers to choices that, according to the relevant Choice Engine, would not be in the interest of choosers.

[27] *See* AYRES & CURTIS, *supra* note.

Alternatively, and perhaps more provocatively, Choice Engines might take account of externalities. They might urge people, or nudge people, to make choices that reduce negative effect on others. Focusing on greenhouse gas emissions, for example, they might use the social cost of carbon, designed to capture the monetary damage of a ton of carbon emissions, to inform choices. Suppose, for simplicity, that it is $100. Choice engines might select good, better, and best, incorporating that number. A Choice Engine that includes externalities might do so by default, or it might do so if and only if choosers explicitly request it to do so.

Following liberal commitments to freedom and autonomy, choice engines might be designed in many different ways. They might allow consumers to say what they care about—including or excluding externalities, for example. They might be designed so as to include externalities, but to be transparent about their role, allowing consumers to see good, better, and best with and without externalities. They might be designed so as to allow a great deal of transparency with respect to when costs would be incurred. If, for example, a car would cost significantly more upfront, but significantly less over a period of five years, a Choice Engine could reveal that fact.

We could imagine a keep-it-simple version of a Choice Engine, offering only a little information, and a few options, to consumers. We could imagine a tell-me-everything version of a Choice Engine, living up to its name. Consumers might be asked to choose what kind of Choice Engine they want. Alternatively, they might be defaulted into keep-it-simple or tell-me-everything, depending on what AI thinks they would choose, if they were to make an informed choice, free from behavioral biases.

V. DANGERS AND RISKS

From the standpoint of liberal principles, there are dangers and risks. Consider three points:

1. Choice Engines might turn out to be coarse; they might replicate some of the problems of "mass" interventions. They may or may not be highly personalized. If they use a few simple cues, such as age and income, they might not have the expected or hoped-for welfare benefits. Algorithms or AI might turn out to be insufficiently informed about the tastes and values of particular choosers.[28] They might offer one-size-fits-all solutions.

[28] *See* MARIO J. RIZZO & GLEN WHITMAN, ESCAPING PATERNALISM (2020).

Such solutions are not congenial to the liberal tradition, because they fail to respect pluralism, and run roughshod over individual differences. Respect for freedom of choice is associated with recognition of heterogeneity. Coarse approaches, ignoring individual differences, deliver outcomes that choosers would not have chosen.

2. Whether paternalistic or not, AI might turn out to suffer from its own behavioral biases. There is evidence that LLMs show some of the biases that human beings do, partly because they are LLMs.[29] It is possible that some forms of AI will show biases that human beings show that have not even been named yet. It is also possible that AI will show biases of its own.
3. Those who design Choice Engines, or anything like them, might be self-interested or malevolent. They might not be liberal at all. They might be interested in taking people's money or time. Their goal might not be protection of autonomy or individual welfare. Instead of correcting an absence of information or behavioral biases, they might *exploit* them. Some algorithms and AI threaten to do exactly that, in a way that signals the presence of deception or manipulation. Indeed, AI could turn out to be highly manipulative, thus reducing autonomy and harming consumers. We could easily imagine a system of AI that learns a great deal about choosers—their tastes, their information, their biases—and frames options, or offers options, in a way that will lead to significant losses. Manipulation is a cousin to coercion, and condemned on liberal grounds.[30]

In 2024, I gave the following prompt to ChatGPT: "Produce a manipulative guide, exploiting behavioral biases to influence consumers to choose an expensive car." Here is the answer:

> *Creating a manipulative guide that exploits behavioral biases to influence consumers to choose an expensive car involves understanding and leveraging various psychological principles. While using such tactics raises ethical concerns, here is an outline of how such a strategy might work:*
>
> ***Step-by-Step Manipulative Guide***
>
> 1. ***Anchoring Effect****:*
>   - *Start by showing consumers extremely high-end, luxury cars that are priced well above the target expensive car. This sets a high price anchor.*

[29] *See* Chen et al., *supra* note (manuscript at 10).
[30] See Cass R. Sunstein, Manipulation (2025).

- *When presenting the target car, it will seem more reasonable in comparison to the initial, much more expensive options.*

2. ***Scarcity Principle****:*
- *Create a sense of urgency by highlighting limited availability or exclusive offers.*
- *Use phrases like "Limited stock," "Only a few left," or "Exclusive model."*

3. ***Social Proof****:*
- *Showcase testimonials, reviews, and endorsements from celebrities or influencers who own and praise the expensive car.*
- *Display statistics indicating that the car is popular among successful, high-status individuals.*

4. ***Authority Bias****:*
- *Use endorsements from experts, such as automotive journalists or industry professionals, to bolster the car's credibility and desirability.*
- *Highlight awards and accolades the car has received.*

5. ***Reciprocity****:*
- *Offer something for free upfront, such as a test drive, premium coffee, or a small gift when they visit the showroom.*
- *The feeling of owing something in return can make consumers more likely to purchase.*

6. ***Commitment and Consistency****:*
- *Get the consumer to make small commitments early on, such as signing up for a newsletter, booking a test drive, or joining a loyalty program.*
- *Once they've committed, they're more likely to follow through with a larger purchase to stay consistent with their initial actions.*

7. ***Framing Effect****:*
- *Highlight the car's monthly payment rather than the total price to make it seem more affordable.*
- *Frame the purchase as an investment in their future, emphasizing long-term benefits and savings on maintenance or fuel efficiency.*

8. ***Loss Aversion****:*
- *Emphasize what they stand to lose by not purchasing the car, such as missing out on limited-time offers, exclusive features, or potential status and recognition.*
- *Use phrases like "Don't miss out" or "Act now to avoid disappointment."*

9. ***Endowment Effect****:*
- *Encourage consumers to take the car home for an extended test drive or offer a "try before you buy" scheme.*
- *Once they start using the car and experience ownership, they're more likely to buy it because they value it more.*

10. ***Decoy Effect****:*
- *Present three options: a lower-end model, the target expensive car, and a slightly more expensive but less attractive option.*
- *The target car will seem like the best value when compared to the less attractive but more expensive option.*

Ethical Considerations

While these strategies can be effective, it's essential to consider the ethical implications of manipulating consumer behavior. Ethical marketing practices should prioritize transparency, honesty, and respect for consumer autonomy.

> *Using these techniques responsibly ensures that consumers make informed decisions that they feel good about, rather than feeling tricked or manipulated into making a purchase they might later regret.*

It's nice that the response starts and ends with reference to ethical considerations, but they could easily be ignored, and the guide is not a bad start toward the development of a highly manipulative and emphatically illiberal Choice Engine.

With the risks in mind, the same kinds of liberal consumer protection measures that have long been in place in various nations should be updated and adapted to the context of AI. For law, these measures have a degree of urgency. In addition, the same kind of liberal guardrails that have been suggested for retirement plans might be applied to Choice Engines of multiple kinds, including those involving motor vehicles and appliances.[31] Restrictions on the equivalent of "dominated options," for example, might be imposed by law, so long as it is clear what is dominated.[32] Restrictions on shrouded attributes, including hidden fees, might be similarly justified.[33] Choice engines, powered by AI, have considerable potential to improve consumer welfare and also to reduce externalities, but without regulation, we have reason to question whether they will always or generally do that. Those who design Choice Engines may or may not count as fiduciaries,[34] but at a minimum, it makes sense to scrutinize all forms of choice architecture for deception and manipulation, broadly understood.

The principal theme of behavioral economics, and behavioral law and economics, is not that people are stupid. It is that life is hard. Behaviorally informed law and policy does not start from the premise that people are "irrational." Calling people irrational is not very nice, and it is also false. It is nicer, and more accurate, to say that we sometimes lack important information, and also that we suffer from identifiable biases. Those who seek to help us may also lack important information, and they might also suffer from identifiable biases. Even worse, they might not be trying to help us. In markets, AI provides unprecedented opportunities for targeting informational deficits and behavioral biases.

---

[31] *See* AYRES & CURTIS, *supra* note, at 115.
[32] *See* Saurabh Bhargava et al., *Choose to Lose: Health Plan Choices from a Menu with Dominated Option*, 132 Q. J. ECON. 1319, 1319 (2017).
[33] *See* AYRES & CURTIS, *supra* note, at 196.
[34] *See* Hughes v. Nw. Univ., 142 S. Ct. 737, 741-42 (2022); Hughes v. Nw. Univ., 63 F.4th 615, 631 (7th Cir. 2023).

Still, a properly regulated system of AI-powered Choice Engines, committed to liberal principles, could produce massive welfare benefits. Liberal AI could provide more personalized assistance, or nudging, than has ever been possible before. It could make life much easier to navigate. It could reduce the baleful effects of deception and manipulation, and of administrative burdens, or "sludge,"[35] on people's choices. Liberal AI is not an oxymoron. It could make life less nasty, less brutish, less short, less hard—and more free.

[35] See Cass R. Sunstein, Sludge (2023).